

УДК 111

DOI: 10.17726/phillT.2016.12.2.5

Онтологические различия информации и данных

Тисов Владислав Владимирович,

*аспирант факультета прикладной математики,
информатики и физики*

*ФГБОУ ВО «Вологодский государственный университет»
Вологда, Россия*

RainheartLang@gmail.com

Аннотация. Термины «информация» и «данные» были введены как чисто технические утилитарные, однако по мере успешного применения прочно вошли в употребление в быту, в самых различных научных областях, стали рассматриваться как философские категории. При этом, зачастую, их общая граница употребления стирается и они используются как эквивалентные. В статье рассмотрены история возникновения, основополагающие свойства этих сущностей и математические подходы к пониманию количества информации, показана возникающая противоречивость при попытке объединять понятия информации и данных. Высказан тезис об онтологическом их различии: информация наличествует только в субъективном бытии, а данные – в объективном. Для обоснования первой части этого тезиса приведены примерные методики представления составляющих сознания из наличия которых, в совокупности с описанной демаркацией, делается вывод о том, что информация является субъективным началом, той частицей, на которые раскладывается всё субъективное. Для второй части – указано материальное основание данных, которое обосновывает имеющиеся свойства, а также показано различие между объективным материальным началом и, собственно, данными. Обозначены методы, порождаемые данной концепцией рассмотрения субъективного и объективного.

Ключевые слова: информация, данные, бытие, субъективное, объективное, энтропия, материя.

Ontological differences of information and data

Tisov Vladislav V.,

*graduate student of Faculty of Applied Mathematics,
Informatics and Physics, Vologda state university
Vologda, Russia*

RainheartLang@gmail.com

Abstract. The terms «information» and «data» were introduced as purely engineering and utilitarian, though, it quickly started to be used in common speech, different scientific areas and appeared as a philosophical categories, due to it's potential range of application. Wherein, the border of usage between these two words is erasing, they often mentioned as equals. In this article the history of appearance, fundamental attributes of these entities and formal approaches for understanding of the information's and data's quality nature are described and the following contradiction is shown. There is a thesis about ontological inequality, stating that the information is only in subjective existence, while data's existence has an objective base. For the first half of the this thesis, there are described methods of approximated representation of components of consciousness, that, together with the way of demarcation of original terms, lead us to conclusion: information is fundament of subjective existence, the element, which all the subjective consists of. For the second part, there is a statement about connection between data and matter, and about the negation of data's and material substance's equality. Also, there are methods, spawned by the given view on subjective and objective nature.

Keywords: information, data, existence, subject, objective, entropy, matter.

Слово «информация» постоянно воспроизводится в потоке сведений и знаний, которые окружают современного человека. В том или ином виде это слово фигурирует во множестве различных терминов – информационные технологии, информационная безопасность, средства информации, поиск информации. Часто с этим словом связаны самые передовые технологии, теории, концепции и идеи, однако что же такое информация?

Можно найти различные определения информации:

- 1) *сведения, независимо от формы их представления, воспринимаемые человеком или специальными устройствами как отражение фактов материального мира в процессе коммуникации (ГОСТ 7.0-99);*

- 2) *знания о предметах, фактах, идеях и т.д., которыми могут обмениваться люди в рамках конкретного контекста (ISO/IEC10746-2:1996);*
- 3) *знания относительно фактов, событий, вещей, идей и понятий, которые в определённом контексте имеют конкретный смысл (ISO/IEC2382:2015).*

Определения подобной степени размытости можно было бы считать приемлемыми в случае со многими трудноуловимыми, неконкретными или чисто конвенционально возникшими сущностями. То, что стоит за современным смыслом слова «информации» было нововведением на границе математической и инженерной областей, в так называемой «теории информации», а это то поле знаний, которое старается быть максимально точным и чётким. Это свойство трудно вяжется с теми формулировками, что были приведены ранее. Тем более странным это кажется, если вспомнить, что эта сущность обладает собственной мерой, единицей измерения для этой меры (бит) и, как следствие, методикой подсчёта этой меры.

Впервые в математическом знании это слово встречается в 1928 году в статье «Transmission of Information» журнала «Bell System Technical Journal» за авторством сотрудника «Western Electric Company» Ральфа Хартли. Статья носила чисто технический характер и не разъясняла значения слова «информация», при этом указывая математическую формулу для вычисления количества этой самой неопределённой информации:

$$H = n * \log_2 s \quad (1)$$

где n – количество символов в сообщении, а s – мощность используемого алфавита (количество используемых символов).

Также Хартли разбирает прикладное применение этой меры для телеграфных и телефонных сетей.

Через двадцать лет это слово вновь возникает в статье всё того же журнала. «A Mathematical Theory of Communication» создана Клодом Шенноном, который в то время, в 1948 году был сотрудником «Bell Laboratories». Шэннон явно говорит, что занимается расширением тех теоретических наработок, что были созданы Ральфом Хартли и Гарри Найквистом. Он ввёл более общую формулу, по сравнению с формулой Хартли, вычисления информации:

$$H = -\sum_{i=1}^s p(i) * \log_2 p(i) \quad (2)$$

где s – мощность алфавита, а $p(i)$ – вероятность появления i -го символа.

Строго говоря, Шэннон рассматривал это выражение как представление «информационной энтропии», то есть меры случайности. Сама формула схожа с формулой энтропии термодинамической системы в Н-теореме Людвиг Больцмана:

$$H = -\int_V P * \ln P * dv^3 \quad (3)$$

Соответственно, в случае дискретного множества символов и замены основания логарифма с числа Эйлера на число 2, она преобразуется к виду Шэннона. Если посмотреть внимательнее, то можно заметить, что в случае $p(i) = const$, то есть когда все события равновероятны, что справедливо для тех механизмов передачи сигналов, которые рассматривал Хартли, получаем следующее: (1)

$$-\sum_{i=1}^s p(i) * \log_2 p(i) = -\sum_{i=1}^s \frac{1}{s} = -\log_2 \frac{1}{s} = -\log_2 \frac{1}{s} = \log_2 \left(\frac{1}{s} \right)^{-1} = \log_2 s \quad (4)$$

Таким образом, описание Шэннона есть обобщение формулы Хартли и тот феномен, который Хартли рассматривал как информацию, подозрительно похож на понятия информационной и термодинамической энтропий. Клод Шэннон, хотя во многом и опирался на изыски своего предшественника, интересовался не столько передачей сигнала, сколько построением теоретической основы для криптографических и криптоаналитических алгоритмов. В криптографии энтропия характеризует степень неожиданности рассматриваемой последовательности и позволяет оценивать качество шифрования для алгоритма, при помощи которого получаются эти данные. Также в своей теории он старался дистанцироваться от смысла информации – ссылок и связей с физическими или концептуальными сущностями, что позволило, по словам Уоррена Уивера, коллеги Шэннона, охватить «не только письменную и устную речь, но также музыку, изобразительное искусство, театр, балет и фактически всё поведение человека». В связи с этим «информация Шеннона» обладала некоторыми неожиданными свойствами для информации, которую рассматривал Хартли.

Шэннон, используя свою теорию, оценил информационную ёмкость английского языка и пришёл к тому выводу, что английский язык избыточен (в силу имеющихся взаимосвязей между символами, которые уменьшают случайность сообщения) на 50% и в среднем каждый символ несёт собой примерно 2.3 бита (единица информации). Проведя небольшой эксперимент, он подтвердил своё предположение о том, что человек на подсознательном уровне чувствует эту взаимосвязанность и поэтому обработка текста состоящего из реальных английских слов гораздо проще, чем для произвольных или полупроизвольных символов. Таким образом, при оценке реально носимого символом объёма информации мы опираемся на статистику употребления символа для конкретного регламента общения. Представим, что мы нашли человека с минимальным знанием английского языка и не имеющего никакого понятия о смысловых связях символов латинского алфавита. Такого человека в современном мире найти, конечно, сложно, но если это мешает интуитивному пониманию ситуации – можно заменить язык на произвольный, не столь популярный, с нераспространённой системой письменности. К примеру, в качестве такого языка может выступать корейский. Тогда отдельные символы в словах будут восприниматься им как равновероятные, соответственно, для него каждый символ несёт $\log_2 26 \approx 4.7$ бита (для случая английского языка). То есть из этих размышлений следует, что количество информации – это величина, зависящая от субъекта, который её воспринимает: такое свойство явно противоречит инженерной интуиции, с которой подходил к вопросу информации Хартли. Также это сложно совместить с современными представлениями об информации в технической среде: количество информации при обработке либо передачи зависит лишь от некоторых физических фактов – объём носителя, время передачи цифрового сигнала и так далее.

Нередко в простой речи и некоторых технических и академических текстах слово «информация» заменяется своим квази-синонимом «данные». В английском языке слово «data» стало употребляться именно-таким образом в инженерной сфере для демаркации той информации, что была передана или представлена в некоторой инфраструктуре вычислительных устройств, то есть изначальный мотив употребления этого слова, вероятно, состоял в демаркации технического использования феномена от его спеку-

лятивной составляющей. Примечательно, что когда Андрей Николаевич Колмогоров занимался переводом статей Клода Шэннона, то вместо «информация» он использовал именно «данные». Сегодня понятие данных трактуется несколько более широко:

- *данные – зарегистрированная информация, представление фактов, понятий или инструкций в форме, приемлемой для общения, интерпретации, или обработки человеком или с помощью автоматических средств (ISO/IEC/IEEE24765-2010);*

- *данные – формы представления информации, с которыми имеют дело информационные системы и их пользователи (ISO/IEC10746-2:1996).*

К данным применима схожая методология измерения, что и для информации, но, при этом, данные инвариантны относительно воспринимающего субъекта, так как они возникают только при достижении объективной конвенции о своей структуре, которая и называется регистрацией информации.

Именно необходимость воспринимающего субъекта, таким образом, отличает информацию от данных, она есть некоторая часть субъективного бытия. Но если рассматривать подобное утверждение – нужно подумать над вопросом: а каков критерий определения того, что является и не является информацией в такой экзистенциальной структуре?

Субъективное бытие, отождествляемое с сознанием, включает в себя структуру восприятий: визуального, аудиального, тактильного, вкусового, моторного и других. Восприятие же представляется в виде комплекса ощущений, которые являются независимыми состояниями. Действительно, ощущения нельзя проанализировать и развёрнуто сказать о том, чем именно они различаются. Можно сказать, конечно, что наука даёт нам этот анализ – раскладывает звуковые волны, сигналы, посылаемые рецепторами в мозг, говорит о качественном различии между, например, красным и синим цветами. Но наука никогда не рассматривает сами феноменальные ощущения сознания в качестве своего объекта или предмета, она занимается лишь рассмотрением физических, измеримых извне коррелятов этих ощущений. Различие между двумя ощущениями – исключительно интуитивное, субъект не может охарактеризовать эту разницу иначе, чем как принципиальную. Это более чем отвечает представлениям о самой информации: её составляющие части есть равноправные элементы некоторого пространства. Все

ощущения – есть информация, соответственно, должна быть возможность измерять её количество. Рассмотрим пару примеров:

Визуальные ощущения логически можно смоделировать при помощи двумерной матрицы цветовых состояний. Размер такой матрицы и количество состояний зависит от разрешающей способности глаза и цветовой чувствительности глаза, соответственно. Соответственно количество информации в отдельном визуальном образе по формуле Хартли есть

$$\sum_{i=1}^m \sum_{j=1}^n \log_2 s_{i,j} \quad (5)$$

Где m , n – размеры матрицы, $s_{i,j}$ – количество состояний для соответствующей дискретной единицы зрительного восприятия – ощущения (представляем формулу так из-за предположения, что разные единицы могут быть неравноправны). Однако эта формула действительна для случая, когда все цветовые состояния для одного «пикселя» зрения равновероятны, что скорее всего неверно, так как визуальное восприятие состоит не только из проецирования света на сетчатку глаза, но также из последующей «фильтрации» сигналов мозгом. Это может привести к зависимости вероятности того или иного цветового состояния от ближайших спроецированных световых волн и от тех ощущений, что были восприняты при предыдущей обработке данных с сетчатки. Это ведёт нас к частному случаю формулы Шэннона и, пожалуй, требует некоторых уточнений, но, фактически, для утверждения о том, что все зрительные ощущения – есть информация, это не необходимо.

Аналогичным образом можно рассмотреть слух. Звуковые волны, являющиеся внешним условием возникновения аудиального ощущения, обычно представляются как одномерный аналоговый сигнал, но в феноменальное ощущение он отображается не путём обычной дискретизации. На уровне органа слуха волна раскладывается и мозг получает уже нечто в виде спектрального слепка, то есть разложение волны на меньшие составляющие. Именно из этого следует контр-интуитивное утверждение о том, что два звука одинаковой частоты, громкости и схожего тембра могут восприниматься как принципиально разные звуки. По формуле Хартли оценим, что количество информации в звуковом ощущении не менее:

$$\log_2 \left(\frac{\log_2 (v_{\max} - v_{\min})}{s_v} * \frac{\log (L_{\max} - L_{\min})}{S_L} T \right) \quad (6)$$

Где $v_{\max}$, $v_{\min}$ соответствуют максимальной и минимальной воспринимаемым частотам, s_v – шаг чувствительности частоты, величина зависящая от того насколько субъект чувствителен к изменению частоты, $L_{\max}$, $L_{\min}$ – максимальная и минимальная воспринимаемые громкости, S_L – шаг чувствительности громкости, аналогично таковому значению для частоты, T – мощность различаемых тембров (они образуют классы эквивалентности для множества всех звуков). Дополнительные бинарные логарифмы закрались в числители по счастливой случайности – распознавание громкости и частоты имеет логарифмическую природу на биологическом уровне и никак не связано с логарифмом у Хартли-Шэннона, где основание подобрано исходя из математических свойств числа 2.

Приведённое не является исчерпывающим при обосновании того, что всякое ощущение сводится к информации, но затруднительно было бы утверждать, что какое-то другое ощущение принципиально отличается по своей природе от тех, что считаются для человека ведущими.

На первый взгляд, куда большее затруднение вызывает другая часть субъективного – мышление. Оно есть, вероятно, более первичный бытийный атрибут нежели банальное восприятие, при этом, привести методику для подсчёта информации мышления, если она существует, гораздо сложнее, чем для того, что рассматривалось прежде. Тем не менее, аргументация о сводимости мышления к информации – нетрудна, поскольку первичное и интуитивное понятие информации как раз-таки более подходит именно для мышления, поскольку традиционно под информацией понимают то, что следует за восприятием, возникающий на его основе априорный синтез:

Знания относительно фактов, событий, вещей, идей и понятий, которые в определённом контексте имеют конкретный смысл.

Если всё же заниматься поиском методики подсчёта информации мышления, то её стоит искать в формальной логике, так как она есть лучшая имеющаяся аппроксимация для предмета мышления. При таком подходе подсчёт количества информации в мысли сводится к формализации рассуждений в виде высказывания языка

логики и использованию того метода, что используется для такового у естественных языков (функция от длины предложения, языкового алфавита и статистической вероятности того или иного факта).

Если исходить из озвученных постулатов:

1) информация для своего существования требует наличия бытийного субъекта;

2) всякая составляющая субъективного бытия, суть, информация,

то сама собой напрашивается эквивалентность терминов субъективного бытия и информации, а точнее следовало бы сказать, что *информация* – это атом субъективного бытия. Всё, что находится в сознании, всё, что человек ощущает, мыслит, представляет, знает, есть информация. Информационная природа – это та общность, что объединяет феномены восприятия и мышления. С позиции дуализма или субъективного идеализма можно сказать, что информация и есть идеальная субстанция. В этом смысле мышление есть обособленный тип восприятия, существующий по сходным законам.

При этом, говорить, что *абсолютно всё* есть информация, было бы верным исключительно с позиции уверенного солипсизма, то есть лишь при отрицании существования объективного мира. Мы уже упоминали термин «данные» и говорили о его сходстве с «информацией», не значит ли это, что данные являются коррелятом информации в объективном бытии? Может ли быть, что данные есть атом объективного, как информация для субъективного? Однако если вспомнить, как именно устроены данные, например, хранящиеся на жёстком диске, то легко понять, что они в объективном бытии ни в коем случае не играют такой роли. В этом смысле минимальным кирпичиком является та сущность, что по-разному приближается, в зависимости от используемой физической модели: тело в терминах Ньютона, кварк в Стандартной модели, струна в Теории Струн и так далее. Тогда следовало бы задаться вопросом: «а как данные соотносятся с этим фундаментальным объектом?» Первое, что приходит в голову: данные – это некоторая структура материи. Если же принять во внимание тот факт, что различные материальные области могут представлять те же самые данные, то выходит, что они образуют некие классы эквивалентности: пин-код от банковской карточки, записанный карандашом на листе бумаги, представляет те же самые данные,

что и при записи в байтах жёсткого диска ноутбука. Однако такое суждение было бы заблуждением, поскольку эту эквивалентность задаёт субъект в момент *интерпретации* данных, то есть при преобразовании их в информацию в субъективном.

Данные можно определить как всевозможные, а не какие-то определённые, конструкторы материи. Они окажутся супервентностны на (полностью зависимы от) физике своих составляющих. При таком определении инвариантность качества данных и их количества становятся очевидными – исходя из законов сохранения материи.

Такое определение, правда, может вызвать возражение о том, что это говорит об эквивалентности фундаментального начала материи и данных, что ранее отрицалось. Если любая совокупность материи есть данные, то как это сочетается с упомянутыми взглядами на это определение как термина, созданного в целях технического использования? Когда мы говорим, скажем о передаче голоса в телефонной сети, мы говорим не о всех данных, которые существуют на этой линии, а о некоторой контекстуальной выборке, интересной интерпретатору, расположенному на другом конце этого туннеля. Если говорить об объективных классах эквивалентности данных, то в рамках объективного рассмотрения следовало бы задавать их исходя из другой особенности: различные данные, подлежащие единообразной интерпретации обладают некоторым родством. Неважно, что за книга перед вами лежит – «Метафизика нравственности» Иммануила Канта, школьный учебник по математике или сборник анекдотов – все они содержат такие данные, что процесс их интерпретации сходен, в данном случае, представлен чтением. То же самое можно сказать и о, скажем, представлении аудиозаписей на виниловых пластинках – процесс интерпретации (имеется в виду граммофоном) одинаков вне зависимости от пластинки. Каждый материальный объект, состоящий из N некоторых частиц можно подать на вход интерпретирующей функции $2^N - 1$ различными способами (согласно простому правилу комбинаторики), именно выбор одного из этих способов задаёт конкретный контекст и делает возможным реальное использование свойств данных.

Сущность данных рассмотрена более поверхностно чем, возможно, хотелось бы, но это не критично для основного выдвигаемого тезиса: существует принципиальное терминологическое

различие между информацией и данными, которое имеет онтологическую подоплёку. Информация является основанием субъективного мира и специфичным коррелятом данных, которые, в свою очередь, существуют объективно и представляют собой конструкт из материального фундамента. Первичная неточность употребления этих понятий и последовавшая путаница, имеющая место сегодня в самых разных кругах и источниках, вызваны поставленными целями и прагматической ориентированностью тех лиц, что их впервые ввели для формализации своих теоретических инженерных идей.

Высказанное утверждение также позволяет рассматривать само субъективное бытие в терминах теории информации, анализировать его и давать формальную характеристику: интроспективно, или на основании внешней объективной корреляции, что, впрочем, требует более тщательной работы над выведением имеющихся связей между этими субъективным и объективным феноменами.

Литература:

1. ГОСТ 7.0-99 «Информационно-библиотечная деятельность, библиография. Термины и определения». (GOST 7.0-99 «Informacionno-bibliotecnaja dejatel'nost', bibliografija. Terminy i opredelenija».)
2. ISO/IEC10746-2:2009 «Информационные технологии. Открытые распределённые технологии. Открытая распределённая разработка. Эталонная модель. Часть 2. Основы». (ISO/IEC10746-2:2009 «Informacionnyye tehnologii. Otkrytye raspredeljonnyye tehnologii. Otkrytaja raspredeljonnaja razrabotka. Jetalonnaja model'. Chast' 2. Osnovy».)
3. ISO/IEC2382:2015 «Information technology – Vocabulary».
4. Глик Д. Информация: История, Теория, Поток. М.: АСТ, 2013. – 576 с. (Glik D. Information: History, Theory, Flow M.: AST, 2013. – 576 p.)
5. Hartley R. V.L. Transmission of Information // Bell System Technical Journal. 1928. – P. 535-563.
6. Shannon C. A Mathematical Theory of Communication // Bell System Technical Journal. 1948. – P. 379-423.
7. ISO/IEC/IEEE24765-2010. «Systems and software engineering – Vocabulary».
8. Chalmers D. J. The Conscious Mind: In Search of a Fundamental Theory. New York, Oxford: Oxford University Press, 2002. – 432 p.
9. Penrose R. The Emperor's New Mind. New York, Oxford: Oxford University Press, 1989. – 480 p.